Ann. Inst. Statist. Math.
Vol. 48, No. 3, 577–602 (1996)

ASYMPTOTICALLY EFFICIENT AUTOREGRESSIVE MODEL SELECTION FOR MULTISTEP PREDICTION

R. J. BHANSALI

Department of Statistics and Computational Mathematics, University of Liverpool, Victoria Building, Brownlow Hill, P.O. Box 147, Liverpool L69 3BX, U.K.

(Received March 1, 1994; revised June 5, 1995)

Abstract. A direct method for multistep prediction of a stationary time series involves fitting, by linear regression, a different autoregression for each lead time, h, and to select the order to be fitted, $\tilde{k}_h$, from the data. By contrast, a more usual 'plug-in' method involves the least-squares fitting of an initial k-th order autoregression, with k itself selected by an order selection criterion. A bound for the mean squared error of prediction of the direct method is derived and employed for defining an asymptotically efficient order selection for h-step prediction, $h \geq 1$; the $S_h(k)$ criterion of Shibata (1980) is asymptotically efficient according to this definition. A bound for the mean squared error of prediction of the plug-in method is also derived and used for a comparison of these two alternative methods of multistep prediction. Examples illustrating the results are given.

Key words and phrases: AIC, FPE, order determination, time series.

1. Introduction

There has been much interest recently in the question of using lead-time dependent estimates or models for multistep prediction of a time series. Thus, Kabaila (1981), Stoica and Soderstrom (1984), Weiss (1991) and Tiao and Xu (1993) consider estimation of the parameters of a specified model separately for each lead time by minimising the sum of squares of the within sample prediction errors. Also, Findley (1983), Gersch and Kitagawa (1983) and Lin and Granger (1994) recommend selecting and fitting a new model for each lead time. Additional references include Greco *et al.* (1984), Milanese and Tempo (1985), del Pino and Marshall (1986), Hurvich (1987) and Wall and Correia (1989). Grenander and Rosenblatt (1954) and Cox (1961) are two early references in which the h-step prediction constants, $h > 1$, for a finite predictor are obtained separately for each h by minimising a relevant mean squared error of prediction. Also, Pillai *et al.* (1992) address the maximum entropy question when the objective is multistep prediction.

From a statistical methodological point of view, the 'direct' method of 'tuning' the parameter estimates or model selection for multistep prediction is quite a general one and it may be applied whenever the notion of a 'true' model generating a time series can be regarded as being implausible and the model fitted for one-step prediction can be viewed simply as a useful approximation for the stochastic structure generating the time series. It is thus probably unsurprising to find that the cited references cover a rather broad range of different classes of time series.

The direct method may be viewed as an alternative to the widely-used 'plug-in' method, see Box and Jenkins (1970), in which the multistep forecasts are obtained from an initial model fitted to the time series by repeatedly using the model with the unknown future values replaced by their own forecasts. Whittle ((1963), p. 36) observed that the plug-in method is optimal in a least-squares sense if the fitted model coincides with that generating the time series, or, in a somewhat restricted sense, for prediction only one step ahead. As, in practice, all fitted models may be incorrect, this observation suggests that for multistep prediction the plug-in method may be improved upon by adopting a different approach and the direct method may be interpreted as being one such approach.

In this paper, a mathematical optimality result for the direct method is established. A result of this type has so far not been available and it contributes towards theoretically underpinning the method in support of which strong empirical evidence has already been presented in the cited references.

Our result is in the context of the autoregressive model fitting approach which provides a natural setting for considering this method because an autoregressive representation exists under weak conditions (Brillinger (1975), p. 78) and the notion of an incorrect model is readily formulated.

On the assumption that T consecutive observations from a discrete-time infinite order autoregressive process, $\{x_t\}$, are available, a new autoregressive model is selected and fitted for each lead time $h \geq 1$. A direct procedure, involving a linear least-squares regression of x_{t+h} on $x_t, x_{t-1}, \ldots, x_{t-k+1}$, is employed for estimating the autoregressive coefficients. Here, $k = \tilde{k}_h$, say, is a random variable and its value is selected anew for each h. In Section 4, an asymptotic lower bound for the h-step mean squared error of prediction of the direct method is derived, and in Section 5 we show that this bound is attained in the limit if $\tilde{k}_h$ is selected by the $S_h(k)$ criterion of Shibata (1980) and also if it is selected by suitable versions of the FPE and AIC criteria of Akaike (1970, 1973).

An asymptotic lower bound for the h-step mean squared error of prediction, $h \geq 1$, of the plug-in method when the initial autoregressive order, $k = \tilde{k}_1$, is selected from the data is also derived in Section 6. The results here point to a two-fold advantage in using the direct method: first, the bound on its h-step mean squared error is smaller than that for the plug-in method, and, secondly, the latter bound is not attainable and thus the actual mean squared error for the plug-in method could be larger than this bound. The difference between these two bounds depends upon h and on the autoregressive coefficients generating $\{x_t\}$. We discuss this point further in Section 7.

In Sections 4 and 5, we also generalise to $h > 1$ the results for $h = 1$ of Shibata (1980), who stated, without proof, that an optimality property established by him

for the $S_1(k)$ criterion would also hold for the $S_h(k)$ criterion when predicting h steps ahead. The generalization involved is non-trivial however, especially a formulation of optimality for the direct method with respect to the plug-in method. The technical machinery required is also more complex because for $h > 1$ an aim of model selection is not to 'whiten' the observed time series so as to produce as residuals a series of (approximately) uncorrelated random variables, but instead it is to produce a series which (approximately) follows a moving average process of order $h - 1$. Furthermore, we also relax a Gaussian assumption invoked by Shibata (1980, 1981). Indeed, the work of this author serves to clarify the relationship between the objectives of consistency and efficiency in order selection by showing that whereas for a finite order process, AIC is not consistent, for an infinite order autoregression it is asymptotically efficient for one-step prediction and for spectral estimation. By contrast, an objective of this paper is to show that this last result does not hold for multistep prediction but that an asymptotically efficient order selection is still possible by selecting a new order for each h with a criterion specifically suggested for this purpose.

2. Preliminaries

Our results hold under a varying combinations of assumptions, and it will be useful to label some of these at the outset.

Suppose that the observed time series is a realization of a stationary process $\{x_t\}$ $(t = 0, \pm 1, \ldots)$ satisfying the following assumptions:

ASSUMPTION 1. That $\{x_t\}$ possesses a one-sided infinite autoregressive representation

$$\sum_{j=0}^{\infty} a(j)x_{t-j} = \epsilon_t, \qquad a(0) = 1,$$

in which $\{\epsilon_t\}$ is a sequence of independent identically distributed random variables, each with mean 0, variance σ^2, and the $a(j)$ are absolutely summable real coefficients such that the polynomial

$$A(z) = \sum_{j=0}^{\infty} a(j)z^j \neq 0, \qquad |z| \leq 1. \tag{2.1}$$

ASSUMPTION 2. That $\{x_t\}$ does not degenerate to a finite order autoregressive process.

The existence of higher order moments of $\{\epsilon_t\}$ is also assumed, but the number of moments varies in different results, to ensure flexibility we state the following assumption:

ASSUMPTION 3. For a value of $s = s_o$, say, to be further specified

$$E(|\epsilon_t|^s) < \infty. \tag{2.2}$$

Our main Theorems 4.1, 5.1 and 6.1 hold when $s_o = 16$ in Assumption 3, but Proposition 3.1 only with $s_o = 8$ and Lemmas 4.1–4.3 with $s_o = 4$.

Assumption 1 ensures that $\{x_t\}$ has a representation

$$x_t = \sum_{j=0}^{\infty} b(j)\epsilon_{t-j}, \qquad b(0) = 1, \tag{2.3}$$

in which the $b(j)$ are absolutely summable and satisfy (2.1) but with $b(j)$ replacing $a(j)$, and $1 + b(1)z + b(2)z^2 + \cdots = [A(z)]^{-1}$.

The covariance function of $\{x_t\}$ is defined by $R(u) = E(x_t x_{t+u})$ $(t, u = 0, \pm 1, \ldots)$, its spectral density function by

$$f(\mu) = (2\pi)^{-1} \sum_{s=-\infty}^{\infty} R(s)\exp(-is\mu), \tag{2.4}$$

and the autoregressive transfer function by $A(\mu) = A\{\exp(-i\mu)\}$.

For any integer n and all $h \geq 1$, we may write

$$x_{n+h} = -\sum_{j=1}^{\infty} \phi_h(j) x_{n+1-j} + z_{n+h}, \tag{2.5}$$

where $-\phi_h(j)$ is the coefficient of x_{n+1-j} $(j = 1, 2, \ldots)$ in the linear least-squares predictor, $\bar{x}_n(h)$, say, of x_{n+h} based on the infinite past, $\{x_t, t \leq n\}$ and

$$z_{n+h} = \sum_{j=0}^{h-1} b(j)\epsilon_{n+h-j} \tag{2.6}$$

is the h-step prediction error. As in Bhansali (1993),

$$\phi_h(j) = -\sum_{u=0}^{j-1} a(u)b(h-1+j-u) \qquad (j = 1, 2, \ldots). \tag{2.7}$$

The corresponding mean squared error of prediction is given by

$$V(h) = E[\{x_{n+h} - \bar{x}_n(h)\}^2] = \sigma^2 \sum_{j=0}^{h-1} b^2(j) \qquad (h \geq 1). \tag{2.8}$$

Denote the k-th order (direct) linear least-squares predictor of x_{n+h} based on the finite past, $\{x_n, x_{n-1}, \ldots, x_{n-k+1}\}$, $k \geq 1$, by

$$\bar{x}_{Dhk}(n) = -\sum_{j=1}^{k} \phi_{Dhk}(j) x_{n+1-j} \tag{2.9}$$

and the corresponding prediction error of $\bar{x}_{Dhk}(n)$ by

$$z_{Dhk}(n) = x_{n+h} - \bar{x}_{Dhk}(n). \tag{2.10}$$

Let $\boldsymbol{\Phi}_{Dh}(k) = [\phi_{Dhk}(1), \ldots, \phi_{Dhk}(h)]'$, $\boldsymbol{\alpha}_h(k) = [R(h), R(h+1), \ldots, R(h+k-1)]'$, $\boldsymbol{\Phi}_h = [\phi_h(1), \phi_h(2), \ldots]'$, $\boldsymbol{R}(k) = [R(u-v)](u, v = 1, \ldots, k)$, $\boldsymbol{R} = [R(u-v)](u, v = 1, 2, \ldots)$. We have,

$$\boldsymbol{\Phi}_{Dh}(k) = -\boldsymbol{R}(k)^{-1}\boldsymbol{\alpha}_h(k), \tag{2.11}$$

$$V_D(h,k) = E[\{z_{Dhk}(n)\}^2] = R(0) + \sum_{j=1}^{k} \phi_{Dhk}(j)R(j+h-1). \tag{2.12}$$

A consequence of Baxter's (1963) inequality is, Bhansali (1993),

$$\sum_{j=1}^{k} |\phi_{Dhk}(j) - \phi_h(j)| \leq \sum_{j=k+1}^{\infty} |\phi_h(j)|, \tag{2.13}$$

and, as $k \to \infty$, $V_D(h,k) \to V(h)$, for each fixed $h \geq 1$.

Having observed $x_1, \ldots, x_T$, suppose as in Shibata (1980) that the k-th order estimate, $\hat{\boldsymbol{\Phi}}_{Dh}(k) = [\hat{\phi}_{Dhk}(1), \ldots, \hat{\phi}_{Dhk}(k)]'$ of $\boldsymbol{\Phi}_{Dh}(k)$ is obtained by regressing x_{t+h} on $x_t, x_{t-1}, \ldots, x_{t-k+1}$ and by minimising the quantity

$$N^{-1} \sum_{t=K_T}^{T-h} \left\{ x_{t+h} + \sum_{j=1}^{k} \phi_{Dhk}(j) x_{t+1-j} \right\}^2 \tag{2.14}$$

with respect to the $\phi_{Dhk}(j)$. Here, K_T is a preassigned upper bound for the autoregressive order, k, $N = T - h - K_T + 1$ and the subscript D stands for the direct method, that is, the $\phi_{Dhk}(j)$ and $\hat{\phi}_{Dhk}(j)$ are obtained in accordance with the direct method.

We make the following assumption regarding K_T:

ASSUMPTION 4. That $\{K_T\}$ is a sequence of positive integers such that as $T \to \infty$, $K_T \to \infty$ but $K_T^2/T \to 0$.

Set $\boldsymbol{X}_t(k) = [x_t, x_{t-1}, \ldots, x_{t-k+1}]'$,

$$\hat{\boldsymbol{R}}_h(k) = N^{-1} \sum_{t=K_T}^{T-h} \boldsymbol{X}_t(k)\boldsymbol{X}_t(k)',$$

$$\hat{\boldsymbol{\alpha}}_h(k) = N^{-1} \sum_{t=K_T}^{T-h} \boldsymbol{X}_t(k) x_{t+h},$$

$$\hat{d}_h(0) = N^{-1} \sum_{t=K_T}^{T-h} x_{t+h}^2,$$

and denote the minimised value of the sum of squares (2.14) by $\hat{V}_{Dh}(k)$, which, as in Bhansali (1993), yields a direct k-th order estimator of $V(h)$. We have,

$$\hat{\boldsymbol{\Phi}}_{Dh}(k) = -\hat{\boldsymbol{R}}_h(k)^{-1}\hat{\boldsymbol{\alpha}}_h(k), \tag{2.15}$$

$$\hat{V}_{Dh}(k) = \hat{d}_h(0) + \hat{\boldsymbol{\alpha}}_h(k)'\hat{\boldsymbol{\Phi}}_{Dh}(k). \tag{2.16}$$

Shibata (1980) proposed that the value of k to be used here be selected by minimising the following criterion:

$$S_h(k) = \hat{V}_{Dh}(k)(N + 2k) \qquad (k = 0, 1, \ldots, K_T). \tag{2.17}$$

We will denote the corresponding order selected by $\hat{k}_{DT}(h)$, i.e.,

$$S_h\{\hat{k}_{DT}(h)\} = \inf_{0 \le k \le K_T} S_h(k).$$

The Cholesky decomposition of $\boldsymbol{R}(k)^{-1}$ is given by

$$\boldsymbol{R}(k)^{-1} = \boldsymbol{G}(k)\boldsymbol{\Sigma}(k)^{-1}\boldsymbol{G}(k)' \qquad (k = 1, 2, \ldots, K_T), \tag{2.18}$$

where $\boldsymbol{G}(k) = [a_{k-v}(u - v)]$, with $a_s(t) = 0$, $t > s$, $a_s(0) = 1$, is lower triangular and $\boldsymbol{\Sigma}(k) = \mathrm{Diag}\{\sigma^2(k-1), \ldots, \sigma^2(0)\}$ is diagonal.

An implication of Assumption 2 is

$$V_D(h, k) \ge V(h), \qquad 1 \le k \le K_T. \tag{2.19}$$

For an infinite-dimensional vector, $\boldsymbol{c} = [c_1, c_2, \ldots]'$ and a positive definite matrix, $\boldsymbol{C}$, we define the norms $\|\boldsymbol{c}\| = (\boldsymbol{c}'\boldsymbol{c})^{1/2}$, $\|\boldsymbol{c}\|_{\boldsymbol{C}} = (\boldsymbol{c}'\boldsymbol{C}\boldsymbol{c})^{1/2}$, $\|\boldsymbol{C}\| = \sup \|\boldsymbol{C}\boldsymbol{c}\|$, $\|\boldsymbol{c}\| \le 1$. Norms for finite-dimensional vectors and matrices are defined analogously and a finite-dimensional vector is sometimes regarded as infinite-dimensional with undefined entries 0.

3. Mean squared error of the direct method

Suppose as in Shibata (1980) and Akaike (1970) that the process to be predicted, $\{y_t\}$, is independent of and has the same stochastic structure as $\{x_t\}$. Let $\bar{y}_{Dhk}(n)$ be defined by (2.9) but with the y's replacing the x's, and let

$$\hat{y}_{Dhk}(n) = -\sum_{s=1}^{k} \hat{\phi}_{Dhk}(s) y_{n+1-s} \tag{3.1}$$

be the corresponding direct estimate of y_{n+h}, where the $\hat{\phi}_{Dhk}(s)$ are based on the observed realization, $x_1, \ldots, x_T$, of $\{x_t\}$ and are calculated as in (2.15).

Let E_y and E_x denote expectations with respect to the $\{y_t\}$ and $\{x_t\}$ processes, respectively, and $E_{y|x}$ the conditional expectation with respect to $\{y_t\}$, conditional on $\{x_t\}$. As in (2.5) and (2.6), we may write, $y_{n+h} = \bar{y}_n(h) + \tilde{z}_{n+h}$, where $\bar{y}_n(h)$ and $\tilde{z}_{n+h}$ are defined analogously to $\bar{x}_n(h)$ and z_{n+h}, respectively, but

with the y's replacing the x's and with $\tilde{z}_{n+h}$ denoting the h-step prediction error of the $\{y_t\}$ process. Hence, on using the identity $E_y = E_x E_{y|x}$, and noting that under the assumption stated at the beginning of this section, $\{y_t\}$ is distributed independently of $\{x_t\}$, the mean squared error of $\hat{y}_{Dhk}(n)$ is given by

$$
\begin{aligned}
(3.2) \qquad & E_y[\{\hat{y}_{Dhk}(n) - y_{n+h}\}^2] \\
& = E_y[\{\hat{y}_{Dhk}(n) - \bar{y}_{Dhk}(n) - \tilde{z}_{n+h}\}^2] \\
& = E_y[\{\tilde{z}_{n+h}\}^2] + E_x E_{y|x}[\{\hat{y}_{Dhk}(n) - \bar{y}_{Dhk}(n)\}^2] \\
& = V(h) + E_x\{\|\hat{\mathbf{\Phi}}_{Dh}(k) - \mathbf{\Phi}_h\|_{\boldsymbol{R}}^2\}.
\end{aligned}
$$

We next develop an asymptotic approximation for the second term to the right of (3.2) by evaluating the probability limit of the quantity occurring inside the curly brackets, since an analytic evaluation of the actual expectation of this term is not straightforward; see Bhansali and Papangelou (1991) for a discussion of this point. To this end, set

$$
(3.3) \qquad Q_D(k) = \|\hat{\mathbf{\Phi}}_{Dh}(k) - \mathbf{\Phi}_h\|_{\boldsymbol{R}}^2.
$$

It follows from (3.2) that the probability limit of $Q_D(k)$ provides an approximation to the mean squared error of h-step prediction up to $V(h)$ when y_{n+h} is predicted by $\hat{y}_{Dhk}(n)$, and, in fact, it could serve as a substitute for this mean squared error.

We may write,

$$
(3.4) \qquad Q_D(k) = \{V_D(h,k) - V(h)\} + \|\hat{\mathbf{\Phi}}_{Dh}(k) - \mathbf{\Phi}_{Dh}(k)\|_{\boldsymbol{R}(k)}^2.
$$

The probability limit of the stochastic term on the right of (3.4) is evaluated below in Proposition 3.1; an explicit proof of this proposition is omitted, but one can be constructed from the results of Section 4.

PROPOSITION 3.1. *Let $\{k_T\}$ be a sequence of integers such that $1 \le k_T \le K_T$ and $k_T \to \infty$ as $T \to \infty$, and suppose that Assumptions 1, 2, 3 with $s_o = 8$ and 4 hold. Then, for each fixed $h \ge 1$,*

$$
(3.5) \qquad p\lim_{T\to\infty} (N/k_T)\|\hat{\mathbf{\Phi}}_{Dh}(k_T) - \mathbf{\Phi}_{Dh}(k_T)\|_{\boldsymbol{R}(k)}^2 = V(h).
$$

For $k = 1, 2, \ldots, K_T$, let

$$
(3.6) \qquad F_{DhT}(k) = \{V_D(h,k) - V(h)\} + (k/N)V(h).
$$

We have the following corollary.

COROLLARY 3.1. *Under the assumptions stated in Proposition* 3.1,

$$
p\lim_{T\to\infty} \{\|\hat{\mathbf{\Phi}}_{Dh}(k_T) - \mathbf{\Phi}_h\|_{\boldsymbol{R}}^2 / F_{DhT}(k_T)\} = 1.
$$

An implication of Corollary 3.1 is that if for a fixed $h \geq 1$, $k = k_T \to \infty$, as $T \to \infty$, then the quantity, $Q_D(k)$, defined by (3.3) and whose probability limit we seek to evaluate, behaves like $F_{DhT}(k)$ asymptotically.

The second term to the right of (3.6) provides a measure of the variance of $\hat{\boldsymbol{\Phi}}_{Dh}(k)$ in estimating $\boldsymbol{\Phi}_{Dh}(k)$, and, by contrast, the first term measures the bias due to employing a finite k-th order predictor for h-step prediction in place of that based on the infinite past.

We introduce the following definition, which is partially motivated by the ideas of Akaike (1970) and Shibata (1980).

DEFINITION 3.1. Let $\{k^*_{DT}(h)\}$ be a sequence of positive integers which attains the minimum of $F_{DhT}(k)$ for each T and a fixed $h \geq 1$, that is,

$$k^*_{DT}(h) = \underset{1 \leq k \leq K_T}{\operatorname{argmin}} F_{DhT}(k).$$

As in Shibata (1980), it follows that if as $T \to \infty$, $K_T \to \infty$ and $K_T/T \to 0$, $F_{DhT}(K_T) \to 0$, and hence for any fixed $h \geq 1$, as $T \to \infty$, $F_{DhT}\{K^*_{DT}(h)\} \to 0$ and $k^*_{DT}(h) \to \infty$.

The following theorem may be established as in Theorem 3.1 of Shibata (1980). It shows that, for each $h \geq 1$, the sequence $k^*_{DT}(h)$ asymptotically yields the value of k at which $Q_D(k)$ is minimised. However, $k^*_{DT}(h)$ is a function of the unknown parameters $V(h)$ and Φ_h and its value will be unknown in practice. In Section 4, Theorem 3.1 is extended to the case where $k = \tilde{k}_{DT}(h)$ is a random variable and the direct method is used for selecting its value separately for each h.

THEOREM 3.1. *Let Assumptions* 1, 3 *with* $s_o = 8$ *and* 4 *hold. Then, for any* $\{k_T\}$ *such that* $1 \leq k_T \leq K_T$, *any* $\delta > 0$, *and each* $h \geq 1$,

$$\lim_{T \to \infty} P([\|\hat{\boldsymbol{\Phi}}_{Dh}(k_T) - \boldsymbol{\Phi}_h\|^2_{\boldsymbol{R}} / F_{DhT}\{k^*_{DT}(h)\}] \geq 1 - \delta) = 1.$$

4. Bound on the mean squared error of prediction of the direct method

From (2.10) and (2.15), we may write, for every $h \geq 1$,

$$\hat{\boldsymbol{\Phi}}_{Dh}(k) - \boldsymbol{\Phi}_D(k) = -\hat{\boldsymbol{R}}_h(k)^{-1} \left\{ N^{-1} \sum_{t=K_T}^{T-h} \boldsymbol{X}_t(k) z_{Dhk}(t) \right\}. \tag{4.1}$$

We show first that $\hat{\boldsymbol{R}}_h(k)^{-1}$ on the right of (4.1) may be replaced by $\boldsymbol{R}(k)^{-1}$. We have the following lemma in which

$$Q_{D1}(k) = \left\| N^{-1} \sum_{t=K_T}^{T-h} \boldsymbol{X}_t(k) z_{Dhk}(t) \right\|^2_{\boldsymbol{R}(k)^{-1}}. \tag{4.2}$$

LEMMA 4.1. *If Assumptions* 1, 3 *with* $s_o = 4$ *and* 4 *hold, for any fixed* $h \geq 1$,

$$p \lim_{T\to\infty} \left\{ \max_{1\leq k\leq K_T} |[\|\hat{\boldsymbol{\Phi}}_{Dh}(k) - \boldsymbol{\Phi}_h\|_{\boldsymbol{R}}^2 / F_{DhT}(k)] - 1| \right\}$$
$$= p \lim_{T\to\infty} \left(\max_{1\leq k\leq K_T} |[Q_{D1}(k) - \{kV(h)/N\}]/F_{DhT}(k)| \right).$$

PROOF. The result follows as in Lemma 3.3 of Shibata (1980), after using (4.1), (3.6) and noting that for all $1 \leq k \leq K_T$,

$$\{F_{DhT}(k)\}^{-1} \leq M(N/k), \tag{4.3}$$

where M denotes a bounded constant, and in the sequel, not necessarily the same constant each time. □

We show next that $z_{Dhk}(t)$ on the right of (4.2) may be replaced by z_{t+h}, that is, $Q_{D1}(k)$ by $Q_{D2}(k)$, where

$$Q_{D2}(k) = \left\| N^{-1} \sum_{t=K_T}^{T-h} \boldsymbol{X}_t(k) z_{t+h} \right\|_{\boldsymbol{R}(k)^{-1}}^2. \tag{4.4}$$

LEMMA 4.2. *Let Assumption* 1 *hold. Then, for all* $1 \leq k \leq K_T$, *and any fixed* $h \geq 1$,

$$N\, E\left[\left\| N^{-1} \sum_{t=K_T}^{T-h} \boldsymbol{X}_t(k)\{z_{Dhk}(t) - z_{t+h}\} \right\|^2 \right]$$
$$\leq M\, k \left\{ \sum_{j=k+1}^{\infty} |\phi_h(j)| \right\}^2.$$

PROOF. The lemma follows from Anderson ((1971), p. 450), see also Lemma 3.1 of Bhansali (1981), and the inequality (2.13). □

LEMMA 4.3. *If Assumptions* 1, 2, 3 *with* $s_o = 4$ *and* 4 *hold, for any fixed* $h \geq 1$,

$$p \lim_{T\to\infty} \left[\max_{1\leq k\leq K_T} |\{Q_{D1}(k) - Q_{D2}(k)\}/F_{DhT}(k)| \right] = 0.$$

PROOF. We have, on using (2.19)

$$\{F_{DhT}(k)\}^{-1} \leq \{V_D(h,k) - V(h)\}^{-1} < M, \qquad 1 \leq k \leq K_T. \tag{4.5}$$

It thus follows from Lemma 4.2 that, with $K = K_T$,

$$E\left[\max_{1\le k\le K_T}\left\|\left\|N^{-1}\sum_{t=K_T}^{T-h}\boldsymbol{X}_t(k)\{z_{Dhk}(t)-z_{t+h}\}\right\|^2/F_{DhT}(k)\right|\right]$$
$$\le M\sum_{k=1}^{K}E\left[\left\|N^{-1}\sum_{t=K_T}^{T-h}\boldsymbol{X}_t(k)\{z_{Dhk}(t)-z_{t+h}\}\right\|^2\right]\le MK_T^2/T$$

and converges to 0 as $T\to\infty$. The lemma may now be established by using the Cauchy-Schwarz inequality and noting that

$$\max_{1\le k\le K_T}\|\boldsymbol{R}(k)^{-1}\|\le M. \tag{4.6}$$ □

Now, consider $Q_{D2}(k)$. On using the Cholesky decomposition, (2.18), for $\boldsymbol{R}(k)^{-1}$, we get

$$Q_{D2}(k)=\sum_{u=1}^{k}\left\{N^{-1}\sum_{t=K_T}^{T-h}z_{t+h}z_{D1k}(t-u+1)\right\}^2\{\sigma^2(k-u)\}^{-1}. \tag{4.7}$$

We show that $Q_{D2}(k)$ may be further simplified by replacing $z_{D1k}(t-u+1)$ and $\sigma^{-2}(k-u)$ by ϵ_{t+1-u} and σ^{-2} respectively. Put

$$Q_{D3}(k)=\sum_{u=1}^{k}\sigma^{-2}\left\{N^{-1}\sum_{t=K_T}^{T-h}z_{t+h}\epsilon_{t+1-u}\right\}^2. \tag{4.8}$$

LEMMA 4.4. *If Assumptions* 1 *and* 3 *with* $s_o = 8$ *hold, for any fixed* $h \ge 1$,

$$N^2E\{|Q_{D2}(k)-Q_{D3}(k)|^2\}\le M\left\{\sum_{u=1}^{k}\sum_{v=k-u+1}^{\infty}|a(v)|\right\}^2\le Mk^2.$$

PROOF. The result follows from (2.13) and the results of Brillinger ((1975), p. 20); see also Lemma 3.2 of Bhansali (1981). □

LEMMA 4.5. *If Assumptions* 1, 2, 3 *with* $s_o = 8$, *and* 4 *hold, for any fixed* $h \ge 1$,

$$p\lim_{T\to\infty}\left[\max_{1\le k\le K_T}|\{Q_{D2}(k)-Q_{D3}(k)\}/F_{DhT}(k)|\right]=0.$$

PROOF. The lemma follows from (4.5) and Lemma 4.4 on noting that

$$\sum_{k=1}^{K}[E|\{Q_{D2}(k)-Q_{D3}(k)\}|^2]^{1/2}\le MK_T^2/T$$

and converges to 0 as $T \to \infty$ by Assumption 4. □

Let $\{z_t\}$ be defined by (2.6), but with t replacing $n+h$ and let $R_{h-1}(u) = E(z_t z_{t+u})$ $(t, u = 0, \pm 1, \ldots)$ denote its covariance function, where

$$R_{h-1}(u) = \sigma^2 \sum_{j=0}^{h-1-|u|} b(j)b(j+|u|) \qquad (|u| \leq h-1), \tag{4.9}$$

and $R_{h-1}(u) = 0$, $|u| \geq h$.

For $h = 1$, the following lemma agrees with Lemmas 3.2 and 3.4 of Shibata (1980), who, however, requires $\{x_t\}$ to be Gaussian.

LEMMA 4.6. *Let Assumptions* 1 *and* 3 *with* $s_o = 16$ *hold. Then, for all* $1 \leq k \leq K_T$, *and every fixed* $h \geq 1$,

$$\begin{aligned}
&(N/k)E\{Q_{D3}(k)\} = V(h), \\
&E[\{N\, Q_{D3}(k) - kV(h)\}^2] = 2k \sum_{u=-h+1}^{h-1} \{R_{h-1}(u)\}^2 + O(1) + O(k^2/N), \\
&N^4 cum\{Q_{D3}(k), Q_{D3}(k), Q_{D3}(k), Q_{D3}(k)\} = kC_1 + O(k^4/N), \\
&E[\{NQ_{13}(k) - kV(h)\}^4] = 12k^2C_2 + 48kC_1 + O(k^4/N),
\end{aligned} \tag{4.10}$$

where,

$$|C_1| \leq 48 \left\{ \sum_{u=-h+1}^{h-1} |R_{h-1}(u)| \right\}^4,$$

$$C_2 = \sum_{u=-h+1}^{h-1} \{1 - (|u|/k)\}\{1 - (|u|/N)\}^2 R_{h-1}(u).$$

PROOF. The lemma follows from Assumption 1 and the results of Brillinger ((1975), p. 20); the details are omitted. □

Proposition 4.1 and Theorem 4.1 below generalise to $h > 1$, Proposition 3.2 and Theorem 3.1 of Shibata (1980).

PROPOSITION 4.1. *Let Assumptions* 1, 2, 3 *with* $s_o = 16$, *and* 4 *hold. Then, for every fixed* $h \geq 1$,

$$p \lim_{T \to \infty} \left[\max_{1 \leq k \leq K_T} |\{\|\hat{\boldsymbol{\Phi}}_{Dh}(k) - \boldsymbol{\Phi}_h\|_{\boldsymbol{R}}^2 / F_{DhT}(k)\} - 1| \right] = 0.$$

PROOF. The result follows from Lemmas 4.1, 4.3, 4.5, 4.6 and an argument employed in the result (3.5) of Shibata (1980). □

THEOREM 4.1. *Let Assumptions* 1, 2, 3 *with* $s_o = 16$, *and* 4 *hold. Then, for any random variable,* $\tilde{k}_{DT}(h)$, *possibly dependent on* $x_1, \ldots, x_T$, *every fixed* $h \geq 1$ *and for any* $\delta > 0$,

$$\lim_{T\to\infty} P([\|\hat{\boldsymbol{\Phi}}_{Dh}\{\tilde{k}_{DT}(h)\} - \boldsymbol{\Phi}_h\|^2_{\boldsymbol{R}}/F_{DhT}\{k^*_{DT}(h)\}] \geq 1 - \delta) = 1.$$

PROOF. The result follows from Proposition 4.1 and Definition 3.1. □

Theorem 4.1 provides an extension of Corollary 3.1 to the situation in which, for each fixed $h \geq 1$, the fitted autoregressive order, $k = \tilde{k}_{DT}(h)$, say, is a random variable and its selected value is possibly based on the observations, $x_1, \ldots, x_T$, on $\{x_t\}$. For this situation, the theorem shows that with any order selection, $\tilde{k}_{DT}(h)$, $Q_D\{\tilde{k}_{DT}(h)\}$, which, as discussed below (3.3), is the substitute mean squared error of prediction up to $V(h)$, is never below $F^*_{DhT}\{k_{DT}(h)\}$, in probability. In this sense $F_{DhT}\{k^*_{DT}(h)\}$ provides a lower bound for $Q_D(k)$ when the autoregressive order is selected separately for each h by the direct method. We accordingly define an asymptotically efficient order selection for h-step prediction as follows:

DEFINITION 4.1. For the direct method of fitting autoregression for h-step prediction, $h \geq 1$, an order selection, $\tilde{k}_{DT}(h)$, is said to be asymptotically efficient if

$$p\lim_{T\to\infty}([\|\hat{\boldsymbol{\Phi}}_{Dh}\{\tilde{k}_{DT}(h)\} - \boldsymbol{\Phi}_h\|^2_{\boldsymbol{R}}/F_{DhT}\{k^*_{DT}(h)\}]) = 1.$$

5. Asymptotically efficient h-step order selection

We now show that the autoregressive order selection by the $S_h(k)$ criterion, (2.17), is asymptotically efficient in the sense of Definition 4.1, and that small changes to this criterion do not affect this property, and thus obtain appropriate generalisations to the case $h > 1$ of the corresponding results of Shibata (1980).

For a fixed $k \geq 1$, let

$$\tilde{V}_{Dh}(k) = N^{-1}\sum_{t=K_T}^{T-h}\{z_{Dhk}(t)\}^2. \tag{5.1}$$

We then have

$$\tilde{V}_{Dh}(k) - \hat{V}_{Dh}(k) = \|\hat{\boldsymbol{\Phi}}_{Dh}(k) - \boldsymbol{\Phi}_{Dh}(k)\|^2_{\hat{\boldsymbol{R}}_h(k)}. \tag{5.2}$$

Hence we may write, for all $k = 1, 2, \ldots, K_T$,

$$S_h(k) = N\ F_{DhT}(k) + g_1(k) + g_2(k) + g_3(k), \tag{5.3}$$

where

$$\begin{aligned} g_1(k) &= 2k\{\hat{V}_{Dh}(k) - V(h)\} + NV(h), \\ g_2(k) &= \{k\ V(h) - N\|\hat{\boldsymbol{\Phi}}_{Dh}(k) - \boldsymbol{\Phi}_{Dh}(k)\|^2_{\hat{\boldsymbol{R}}_h(k)}\}, \\ g_3(k) &= N\{\tilde{V}_{Dh}(k) - V_D(h,k)\}. \end{aligned}$$

Proposition 4.1 shows that as compared with the first term, the fourth term to the right of (5.3) is small, uniformly in $1 \leq k \leq K_T$. Lemma 5.1 below examines the second term.

LEMMA 5.1. *If Assumptions* 1, 2, 3 *with* $s_o = 16$, *and* 4 *hold, for any fixed* $h \geq 1$,

$$p \lim_{T \to \infty} \left[\max_{1 \leq k \leq K_T} |k\{\hat{V}_{Dh}(k) - V(h)\}/\{NF_{DhT}(k)\}| \right] = 0.$$

PROOF. On using (5.2), we may write

$$|\hat{V}_{Dh}(k) - V(h)| \leq \|\hat{\boldsymbol{\Phi}}_{Dh}(k) - \boldsymbol{\Phi}_{Dh}(k)\|^2_{\hat{\boldsymbol{R}}_h(k)} + |\tilde{V}_{Dh}(k) - V_D(h,k)| + |V_D(h,k) - V(h)|. \tag{5.4}$$

Now, by Lemma 3.3 of Shibata (1980) and our Proposition 4.1,

$$p \lim_{T \to \infty} \max_{1 \leq k \leq K_T} [k\|\hat{\boldsymbol{\Phi}}_{Dh}(k) - \boldsymbol{\Phi}_{Dh}(k)\|^2_{\hat{\boldsymbol{R}}_h(k)}/\{NF_{DhT}(k)\}] = 0. \tag{5.5}$$

Also, since $E[\{\tilde{V}_{Dh}(k) - V_D(h,k)\}^2] = O(N^{-1})$, uniformly in $1 \leq k \leq K_T$,

$$\sum_{k=1}^{K} E[\{\tilde{V}_{Dh}(k) - V_D(h,k)\}^2] \leq MK_T/N$$

and converges to 0 as $T \to \infty$. Thus, in view of (4.5),

$$p \lim_{T \to \infty} \max_{1 \leq k \leq K_T} [k|\tilde{V}_{Dh}(k) - V_D(h,k)|/\{NF_{DhT}(k)\}] = 0. \tag{5.6}$$

The lemma now follows from Assumption 2 and the inequality (4.5). □

Consider now the last term to the right of (5.3). If $\{k^*_{hT}\}$ is a sequence of positive integers such that $k^*_{hT} \to \infty$, as $T \to \infty$, then for every $1 \leq k \leq K_T$ and each fixed $h \geq 1$ we may write

$$\begin{aligned} &[\tilde{V}_{Dh}(k^*_{hT}) - V_D(h,k^*_{hT})] - [\tilde{V}_{Dh}(k) - V_D(h,k)] \\ &\quad = 2[\boldsymbol{\Phi}_{Dh}(k^*_{hT}) - \boldsymbol{\Phi}_{Dh}(k)]'[\hat{\boldsymbol{\alpha}}_h(K_T) - \boldsymbol{\alpha}_h(K_T)] \\ &\qquad + [\boldsymbol{\Phi}_{Dh}(k^*_{hT}) - \boldsymbol{\Phi}_{Dh}(k)]'\{\hat{\boldsymbol{R}}_h(K_T) - \boldsymbol{R}(K_T)\} \\ &\qquad \cdot [\boldsymbol{\Phi}_{Dh}(k^*_{hT}) + \boldsymbol{\Phi}_{Dh}(k)], \end{aligned} \tag{5.7}$$

where $\boldsymbol{\Phi}_{Dh}(k^*_{hT})$ is considered as a $K_T \times 1$ vector with undefined entries set equal to zero. We show below that the contribution of the two terms to the right of (5.7) is asymptotically negligible.

LEMMA 5.2. *Let Assumptions* 1, 2 *and* 4 *hold and that* $k^*_{hT} \to \infty$, *as* $T \to \infty$. *Then, for every fixed* $h \geq 1$, *as* $T \to \infty$,

$$\max_{1 \leq k \leq K_T} |\{\boldsymbol{\Phi}_{Dh}(k^*_{hT}) - \boldsymbol{\Phi}_{Dh}(k)\}'\{\hat{\boldsymbol{\alpha}}_h(K_T) - \boldsymbol{\alpha}_h(K_T)\}/\{F_{DhT}(k)\}|,$$

and

$$\max_{1\le k\le K_T} |\{\boldsymbol{\Phi}_{Dh}(k^*_{hT}) - \boldsymbol{\Phi}_{Dh}(k)\}'\{\hat{\boldsymbol{R}}_h(K_T) - \boldsymbol{R}(K_T)\} \cdot \{\boldsymbol{\Phi}_{Dh}(k^*_{hT}) + \boldsymbol{\Phi}_{Dh}(k)\}/\{F_{DhT}(k)\}|$$

converge in probability to zero.

PROOF. The result follows from (4.5) and the inequality (2.13) by noting that

$$E\{|\hat{d}_h(u) - R(h-1+u)|\} \le MT^{-1/2}, \quad 1 \le u \le K_T,$$
$$E\{|D^{(T)}(u,v) - R(v-u)|\} \le MT^{-1/2}, \quad 1 \le u, v \le K_T,$$

where $\hat{d}_h(u)$ denotes the typical element of $\hat{\boldsymbol{\alpha}}_h(K_T)$ and $D^{(T)}(u,v)$ that of $\hat{\boldsymbol{R}}_h(K_T)$. □

We now establish the asymptotic efficiency of the order selected by the $S_h(k)$ criterion for all fixed $h \ge 1$. We also show that small changes to the $S_h(k)$ criterion do not affect the asymptotic efficiency of the selected order.

Let $k^o_{DT}(h)$ be the order selected by minimising a new criterion,

$$S^o_h(k) = \{N + 2k + \Omega_{hT}(k)\}\hat{V}_{Dh}(k), \quad 1 \le k \le K_T, \tag{5.8}$$

in which $\Omega_{hT}(k)$ denotes some function of k.

The following theorem generalizes Theorems 4.1 and 4.2 of Shibata (1980) to the case $h > 1$:

THEOREM 5.1. *Let Assumptions* 1, 2, 3 *with* $s_o = 16$, *and* 4 *hold, and that* $\Omega_{hT}(k)$ *in* (5.8) *is such that*

i) $p\lim_{T\to\infty}\left\{\max_{1\le k\le K_T} |\Omega_{hT}(k)|/N\right\} = 0;$

ii) $p\lim_{T\to\infty}\left\{\max_{1\le k\le K_T} |[\Omega_{hT}(k) - \Omega_{hT}\{k^*_{DT}(h)\}]/[NF_{DhT}(k)]\right\} = 0.$

Then for every fixed $h \ge 1$,

a) $p\lim_{T\to\infty}[\|\hat{\boldsymbol{\Phi}}_{Dh}\{\hat{k}_{DT}(h)\} - \boldsymbol{\Phi}_h\|^2_{\boldsymbol{R}}/F_{DhT}\{k^*_{DT}(h)\}] = 1;$

b) *the selection,* $k^o_{DT}(h)$, *is asymptotically efficient.*

PROOF. To establish a), observe that for all $T = 1, 2, \ldots,$

$$P[S_h\{\hat{k}_{DT}(h)\} \le S_h\{k^*_{DT}(h)\}] = 1.$$

By (5.3), however, this result implies that

$$P([F_{DhT}\{\hat{k}_{DT}(h)\} - F_{DhT}\{k^*_{DT}(h)\}] \le Z_{hT}) = 1,$$

where

$$Z_{hT} = N^{-1} \sum_{i=1}^{3} \{g_i(k^*_{DT}(h)) - g_i(\hat{k}_{DT}(h))\}.$$

Now, Proposition 4.1 and Lemmas 5.1 and 5.2 imply that, as $T \to \infty$, $Z_{hT} \to 0$, in probability. It follows therefore by a standard argument that for any $\delta > 0$,

$$\lim_{T\to\infty} P([F_{DhT}\{\hat{k}_{DT}(h)\}/F_{DhT}\{k^*_{DT}(h)\}] \leq 1 + \delta) = 1.$$

The result a) now follows from Proposition 4.1 by deducing first that this last result and the definition of $k^*_{DT}(h)$ imply that

$$p \lim_{T\to\infty} [F_{DhT}\{\hat{k}_{1T}(h)\}/F_{DhT}\{k^*_{DT}(h)\}] = 1.$$

Also, b) follows from a), conditions i) and ii) and Proposition 4.1; see also Theorem 4.2 of Shibata (1980). □

As examples of the alternative criteria whose asymptotic efficiency also follows from Theorem 5.1, we have the following:

1) The h-step Final Prediction Error Criterion:

$$FPEh(k) = \hat{V}_{Dh}(k)(1 + k/T)(1 - k/T)^{-1} \quad (k = 0, 1, \ldots, K_T); \tag{5.9}$$

2) The h-step information criterion

$$AICh(k) = T \ln \hat{V}_{Dh}(k) + 2k \quad (k = 0, 1, \ldots, K_T), \tag{5.10}$$

which, for $h = 1$, reduce to the FPE and AIC criteria of Akaike (1970, 1973) respectively; see also Bhansali (1993).

6. Bound for the mean squared error of the plug-in method

Let $\hat{\boldsymbol{a}}(k) = [\hat{a}_k(1), \ldots, \hat{a}_k(k)] = \hat{\boldsymbol{\Phi}}_{D1}(k)$ be the k-th order least-squares estimate of the autoregressive coefficients, and denote the corresponding theoretical parameter by $\boldsymbol{a}(k) = [a_k(1), \ldots, a_k(k)]' = \boldsymbol{\Phi}_{D1}(k)$, $k = 1, \ldots, K_T$. We also write $\sigma^2(k) = V_D(1, k)$.

Now, put

$$\hat{\boldsymbol{\Gamma}}(k) = \begin{bmatrix} -\hat{\boldsymbol{a}}(k)' \\ \cdots\cdots\cdots \\ \boldsymbol{I}_{k-1} \mid \boldsymbol{O}_{k-1} \end{bmatrix}, \tag{6.1}$$

where $\boldsymbol{I}_{k-1}$ and $\boldsymbol{O}_{k-1}$ respectively denote an identity matrix and a vector of zeros of dimension $k - 1$, and define $\boldsymbol{\Gamma}(k)$ analogously but with $\boldsymbol{a}(k)$ replacing $\hat{\boldsymbol{a}}(k)$ in (6.1).

As in Yamamoto (1976) and Lewis and Reinsel (1985), the k-th order plug-in estimator, $\hat{\boldsymbol{\Phi}}_{Ph}(k) = [\hat{\phi}_{Phk}(1), \ldots, \hat{\phi}_{Phk}(k)]'$, of the h-step prediction constants is given by, with $e_k = [1, 0, \ldots, 0]'$,

$$\hat{\boldsymbol{\Phi}}_{Ph}(k) = -\boldsymbol{e}_k'\hat{\boldsymbol{\Gamma}}(k)^h \qquad (h \geq 1), \tag{6.2}$$

the corresponding theoretical parameter, $\boldsymbol{\Phi}_{Ph}(k) = [\phi_{Phk}(1), \ldots, \phi_{Phk}(k)]'$ is defined analogously but with $\boldsymbol{\Gamma}(k)$ replacing $\hat{\boldsymbol{\Gamma}}(k)$ on the right of (6.2).

Denote the k-th order plug-in predictor of x_{n+h}, $h \geq 1$, based on $\{x_n, x_{n-1}, \ldots, x_{n-k+1}\}$ by

$$\bar{x}_{Phk}(n) = -\sum_{j=1}^{k} \phi_{Phk}(j) x_{n+1-j}. \tag{6.3}$$

The corresponding h-step prediction error is denoted by

$$z_{Phk}(n) = x_{n+h} - \bar{x}_{Phk}(n)$$

and the mean squared error of prediction by

$$V_P(h, k) = E[\{z_{Phk}(n)\}^2].$$

We have, as in Bhansali (1981),

$$V_P(h, k) = \int_{-\pi}^{\pi} |B_{h-1,k}(\mu)|^2 |A_k(\mu)|^2 f(\mu) d\mu, \tag{6.4}$$

where, with $\boldsymbol{\beta}_k(j) = \boldsymbol{e}_k'\boldsymbol{\Gamma}(k)^j \boldsymbol{e}_k$ $(j \geq 1)$, $\beta_k(0) = 1$, $B_{h-1,k}(\mu)$ denotes the transfer function of the $\beta_k(j)$ $(j = 0, 1, \ldots, h-1)$ and $A_k(\mu)$ that of the $a_k(j), a_k(0) = 1$.

We note that if $\{x_t\}$ is a finite autoregressive process of order m, $m \leq k$, then, for all $h \geq 1$,

$$\boldsymbol{\Phi}_{Dh}(k) = \boldsymbol{\Phi}_{Ph}(k) = \boldsymbol{\Phi}(k, h),$$

where $\boldsymbol{\Phi}(k, h) = [\phi_h(1), \ldots, \phi_h(k)]'$ and $\phi_h(j) = 0$, $j > m$, and now $\hat{\boldsymbol{\Phi}}_{Ph}(k)$, provides an approximate maximum likelihood estimator of $\boldsymbol{\Phi}(k, h)$, and it is asymptotically efficient for all $h \geq 1$; by contrast, $\hat{\boldsymbol{\Phi}}_{Dh}(k)$ is asymptotically inefficient for $h > 1$. Under Assumption 2, however, $\boldsymbol{\Phi}_{Dh}(k)$ may be different from $\boldsymbol{\Phi}_{Ph}(k)$ and for all $k \geq 1$ and each fixed $h > 1$,

$$V_P(h, k) \geq V_D(h, k) \geq V(h). \tag{6.5}$$

Moreover, since, as $k \to \infty$, $a_k(j) \to a(j)$, for every fixed $j \geq 1$, $\phi_{Phk}(j) \to \phi(j)$ and $V_P(h, k) \to V(h)$, for all $h \geq 1$.

Now, as in Section 3, suppose that the process to be predicted, $\{y_t\}$, is independent of and has the same stochastic structure as $\{x_t\}$. Let $\bar{y}_{Phk}(n)$ be defined by (6.3) but with y_i replacing x_i, $i = 1, \ldots, k$ and let

$$\hat{y}_{Phk}(n) = -\sum_{s=1}^{k} \hat{\phi}_{Phk}(s) y_{n+1-s}$$

be the corresponding plug-in estimate of y_{n+h}, where the $\hat{\phi}_{Phk}(s)$ are based on $x_1, \ldots, x_T$ and are calculated as in (6.2).

As in (3.2), the mean squared error of prediction of $\hat{y}_{Phk}(n)$ is given by

$$E_y[\{\hat{y}_{Phk}(n) - y_{n+h}\}^2] = V(h) + E_x\{Q_P(k)\}$$

where

$$Q_P(k) = \|\hat{\boldsymbol{\Phi}}_{Ph}(k) - \boldsymbol{\Phi}_h\|_{\boldsymbol{R}}^2.$$

Moreover, we now have,

$$Q_P(k) = \{V_P(h,k) - V(h)\} + \|\hat{\boldsymbol{\Phi}}_{Ph}(k) - \boldsymbol{\Phi}_{Ph}(k)\|_{\boldsymbol{R}(k)}^2 + w(k), \tag{6.6}$$

where

$$w(k) = 2\{\boldsymbol{\Phi}_{Ph}(k) - \boldsymbol{\Phi}_h\}'\boldsymbol{R}\{\hat{\boldsymbol{\Phi}}_{Ph}(k) - \boldsymbol{\Phi}_{Ph}(k)\}. \tag{6.7}$$

The probability limits of the stochastic terms on the right of (6.6) and (6.7) may be evaluated as in Section 3 and are given below in Proposition 6.1; for related results see Lewis and Reinsel (1985) and Bhansali (1993). To save space, an explicit proof of this proposition is not given.

PROPOSITION 6.1. *Under the assumptions stated in Proposition* 3.1, *for each fixed* $h \geq 1$,

$$p\lim_{T\to\infty} (N/k_T)\|\hat{\boldsymbol{\Phi}}_{Ph}(k_T) - \boldsymbol{\Phi}_{Ph}(k_T)\|_{\boldsymbol{R}(k)}^2 = V(h), \tag{6.8}$$

$$p\lim_{T\to\infty} (N/k_T)|w(k_T)| = 0. \tag{6.9}$$

Now, for $k = 1, 2, \ldots, K_T$, let

$$F_{PhT}(k) = \{V_P(h,k) - V(h)\} + (k/N)V(h). \tag{6.10}$$

We have the following corollary, which shows that, if the initial fitted order $k = k_T$, of the plug-in method is a function of T such that $k_T \to \infty$, as $T \to \infty$, but sufficiently slowly, then, its substitute mean squared error of prediction up to $V(h)$, $Q_P(k)$, asymptotically behaves like $F_{PhT}(k)$ for each fixed $h \geq 1$.

COROLLARY 6.1. *Under the assumptions stated in Proposition* 6.1,

$$p\lim_{T\to\infty} \{\|\hat{\boldsymbol{\Phi}}_{Ph}(k_T) - \boldsymbol{\Phi}_h\|_{\boldsymbol{R}}^2 / F_{PhT}(k_T)\} = 1.$$

It may be observed that Proposition 6.1 and Corollary 6.1 establish for the plug-in method results parallel to Proposition 3.1 and Corollary 3.1 established earlier in Section 3 for the direct method. Moreover, for the plug-in method (6.10) affords an interpretation similar to that of (3.6): thus, the second term to its

right provides a measure of the variance of $\hat{\mathbf{\Phi}}_{Ph}(k)$ for estimating $\mathbf{\Phi}_{Ph}(k)$ and the first term measures the bias in using a finite k-th order predictor in place of one based on the infinite past. Now, however, an interesting contrast between these two methods emerges. Proposition 6.1 shows that asymptotically the direct and plug-in methods have the same variance, but, by (6.5), the bias of the plug-in method is never smaller. Thus, for all fixed T and h, and each $k \geq 1$,

$$F_{PhT}(k) \geq F_{DhT}(k). \tag{6.11}$$

We now extend this last comparison to the situation in which the initial selected order, k, of the plug-in method is a random variable whose value is possibly selected by an order determining criterion. The basic approach we take for this purpose is to derive a lower bound for the mean squared error of prediction of the plug-in method applicable to this situation and then to compare the magnitude of this bound with that derived earlier in Section 4 for the direct method.

We first introduce the following definition, which parallels Definition 3.1, but applies to the plug-in method.

DEFINITION 6.1. Let $\{k^*_{PT}(h)\}$ be a sequence of positive integers which attains the minimum of $F_{PhT}(k)$ for each T and a fixed $h \geq 1$, that is,

$$k^*_{PT}(h) = \underset{1 \leq k \leq K_T}{\operatorname{argmin}} F_{PhT}(k).$$

Like $k^*_{DT}(h), k^*_{PT}(h) \to \infty$, as $T \to \infty$, since if $K_T \to \infty$ but $K_T/T \to 0$, $F_{PhT}(K_T) \to 0$.

The following theorem establishes for the plug-in method a result similar to that established in Theorem 3.1 for the direct method since it shows that, for each $h \geq 1$, the sequence $k^*_{PT}(h)$ asymptotically yields the values of k that minimises $Q_P(k)$.

THEOREM 6.1. *Let Assumptions* 1, 3 *with* $s_0 = 8$ *and* 4 *hold. Then, for any* $\{k_T\}$ *such that* $1 \leq k_T \leq K_T$, *any* $\delta > 0$, *and each* $h \geq 1$,

$$\lim_{T \to \infty} P([\|\hat{\mathbf{\Phi}}_{Ph}(k_T) - \mathbf{\Phi}_h\|^2_{\boldsymbol{R}}/F_{PhT}\{k^*_{PT}(h)\}] \geq 1 - \delta) = 1.$$

However, $k^*_{PT}(h)$, like $k^*_{DT}(h)$ is unknown in practice as it is a function of the unknown parameters $V(h)$ and $\mathbf{\Phi}_h$. Below, we extend Theorem 6.1 to the situation in which $k = \tilde{k}_{PT}$, say, may be chosen by FPE, AIC or related criteria; the latter include, with $\alpha = 2$, and $h = 1$, the criteria to be introduced in Section 7 by (7.3), and the former two are given by (5.9) and (5.10), but with $h = 1$.

Put $\tilde{N} = T - K_T$,

$$\hat{\boldsymbol{Y}}(k) = (\tilde{N})^{-1} \sum_{t=K_T}^{T-1} z_{D1k}(t+1)\boldsymbol{X}_t(k),$$

and for $u, v = 0, 1, \ldots, h-1$, let

$$\boldsymbol{H}_{uv}(k) = \boldsymbol{\Gamma}(k)^{h-1-u}\boldsymbol{R}(k)\boldsymbol{\Gamma}(k)'^{h-1-v},$$

$$\hat{\boldsymbol{L}}_{uv}(k) = \hat{\boldsymbol{R}}(k)^{-1}\boldsymbol{H}_{uv}(k)\hat{\boldsymbol{R}}(k)^{-1},$$

$$\boldsymbol{W}_{uv}(k) = \hat{\boldsymbol{Y}}(k)'\hat{\boldsymbol{L}}_{uv}(k)\hat{\boldsymbol{Y}}(k), \tag{6.12}$$

$$\boldsymbol{W}_{1uv}(k) = \boldsymbol{Y}(k)'\boldsymbol{L}_{uv}(k)\boldsymbol{Y}(k), \tag{6.13}$$

where $\boldsymbol{Y}(k)$ is obtained from $\hat{\boldsymbol{Y}}(k)$ by replacing $z_{D1k}(t+1)$ with ϵ_{t+1} and $\boldsymbol{L}_{uv}(k)$ from $\hat{\boldsymbol{L}}_{uv}(k)$ by replacing $\hat{\boldsymbol{R}}(k)^{-1}$ with $\boldsymbol{R}(k)^{-1}$, and $\hat{\boldsymbol{R}}(k) = \hat{\boldsymbol{R}}_1(k)$ is as in (2.15) but with $h = 1$.

We may write (Bhansali (1981)),

$$\|\hat{\boldsymbol{\Phi}}_{Ph}(k) - \boldsymbol{\Phi}_{Ph}(k)\|^2_{\boldsymbol{R}(k)} = \sum_{u=0}^{h-1}\sum_{v=0}^{h-1}\hat{\beta}_k(u)\hat{\beta}_k(v)W_{uv}(k), \tag{6.14}$$

where $\hat{\beta}_k(j) = \boldsymbol{e}_k'\hat{\boldsymbol{\Gamma}}(k)^j\boldsymbol{e}_k$ $(j = 1, 2, \ldots, h-1)$ and $\hat{\beta}_k(0) = 1$.

The following lemma shows that for each fixed (u, v) the error in replacing $W_{uv}(k)$ by $W_{1uv}(k)$ is asymptotically negligible.

LEMMA 6.1. *Let Assumptions* 1, 2, 3 *with* $s_o = 4$ *and* 4 *hold. Then, for every fixed* $u, v = 0, 1, \ldots, h-1$ *and each fixed* $h \geq 1$,

$$p\lim_{T\to\infty}\left\{\max_{1\leq k\leq K_T}|[W_{1uv}(k) - W_{uv}(k)]/F_{PhT}(k)|\right\} = 0.$$

PROOF. The argument is in two parts: the first step is to show that $\hat{\boldsymbol{Y}}(k)$ may be replaced by $\boldsymbol{Y}(k)$ on the right of (6.12), and the second to show that $\hat{\boldsymbol{L}}_{uv}(k)$ may also be replaced by $\boldsymbol{L}_{uv}(k)$. To save space the details are omitted, but the method used parallels that given earlier in the proofs of Lemmas 4.1 and 4.2. □

We have, for $u, v = 0, 1, \ldots, h-1$,

$$E\{W_{1uv}(k)\} = \operatorname{tr}\{\sigma^2\boldsymbol{R}(k)^{-1}\boldsymbol{H}_{uv}(k)\} = W_{2uv}(k), \quad \text{say.}$$

The orders of magnitude of the variance and the fourth central moment of $W_{1uv}(k)$ are evaluated in the following lemma.

LEMMA 6.2. *Let Assumptions* 1, 4 *and* 3 *with* $s_o = 16$ *hold. For every* $u, v = 0, 1, \ldots, h-1$, *and each* $h \geq 1$, *uniformly in* $1 \leq k \leq K_T$,

$$E[\{NW_{1uv}(k) - W_{2uv}(k)\}^2] = O(k) + O(k^2/\tilde{N}); \tag{6.15}$$

$$N^4\operatorname{cum}\{W_{1uv}(k), W_{1uv}(k), W_{1uv}(k), W_{1uv}(k)\} = O(k^2) + O(k^4/\tilde{N});$$

$$E[\{NW_{1uv}(k) - W_{2uv}(k)\}^4] = O(k^2) + O(k^4/\tilde{N}).$$

PROOF. The lemma follows by applying the rules for evaluating cumulants given by Brillinger ((1975), p. 19). □

We now show that $W_{1uv}(k)$ may be replaced by $W_{2uv}(k)$.

LEMMA 6.3. *Let Assumptions* 1, 2, 3 *with* $s_o = 16$, *and* 4 *hold. Then, for* $u, v = 0, 1, \ldots, h-1$, *and each fixed* $h \geq 1$,

$$p \lim_{T\to\infty} \left[\max_{1\leq k\leq K_T} |\{W_{1uv}(k) - W_{2uv}(k)\}/F_{PhT}(k)| \right] = 0.$$

PROOF. The lemma follows from Lemma 6.2; see also result (3.5) of Shibata (1980). □

In the following lemma, an expression for $W_{2uv}(k)$ is obtained by using the Cholesky decomposition (2.18) for $\boldsymbol{R}(k)^{-1}$.

LEMMA 6.4. *For every* $u, v = 0, 1, \ldots, h-1$, *each fixed* $h \geq 1$, *and all* $k \geq h$, *with* $s = h-1-u$, $t = h-1-v$, $\bar{k} = k - \max(s,t) - 1$, *as* $k \to \infty$,

$$\sigma^{-2} W_{2uv}(k) = M + \int_{-\pi}^{\pi} \left[\sum_{j=0}^{\bar{k}} \{|A_j(\mu)|^2/\sigma^2(j)\} \right] \exp\{i|s-t|\mu\} f(\mu) d\mu. \tag{6.16}$$

PROOF. We have, on using (2.18), for all u, v

$$\sigma^{-2} W_{2uv}(k) = \operatorname{tr}\{\boldsymbol{G}(k)' \boldsymbol{H}_{uv}(k) \boldsymbol{G}(k) \boldsymbol{\Sigma}(k)^{-1}\}. \tag{6.17}$$

The lemma may now be established by noting that the submatrix in the bottom $(k-s) \times (k-t)$ right hand corner of $\boldsymbol{H}_{uv}(k)$ has $R(p-q)$ as its typical element $(p = 1, \ldots, k-s;\ q = 1, \ldots, k-t)$, and partitioning the matrices $\boldsymbol{G}(k)'$, $\boldsymbol{G}(k)$ and $\boldsymbol{\Sigma}(k)^{-1}$ analogously but ensuring that the dimensions of the relevant submatrices of the matrix product on the right of (6.17) match. □

An implication of Lemma 6.4 is that a uniform bound, valid for all $1 \leq k \leq K_T$, for the mean squared error of the plug-in method may not take a convenient form; a useful approximation may be obtained, however, by restricting the choice of k to the range $k_T \leq k \leq K_T$, where $k_T \to \infty$ as $T \to \infty$ is a sequence of positive integers. This restriction may be justified because the order selected by AIC and related criteria have this property.

LEMMA 6.5. *Let Assumptions* 1 *and* 3 *with* $s_o = 4$ *hold. For* $u, v = 0, 1, \ldots, h-1$, *each fixed* $h \geq 1$ *and every divergent sequence* $\{k_T\}$, *with* $\delta_{uv} = 1$, $u = v$, $\delta_{uv} = 0$, $u \neq v$, *as* $T \to \infty$,

$$|\{W_{2uv}(k) - k\sigma^2\delta_{uv}\}/\{\tilde{N} F_{PhT}(k)\}| \tag{6.18}$$

converges to 0 *uniformly in* $k_T \leq k \leq K_T$.

If also Assumptions 2, 3 *with* $s_o = 16$ *and* 4 *hold,*

$$\text{(6.19)} \qquad \text{i)} \quad p\lim_{T\to\infty} |\hat{\beta}_k(u) - \beta_k(u)| = 0, \text{ uniformly in } 1 \leq k \leq K_T;$$

$$\text{(6.20)} \qquad \text{ii)} \quad |\beta_k(u) - b(u)| \to 0, \text{ uniformly in } k_T \leq k \leq K_T.$$

PROOF. The result (6.18) follows directly from Lemma 6.4, while (6.19) and (6.20) are established recursively in $j = 1, 2, \ldots, u$ on noting that by Proposition 4.1, see also Shibata (1981), $\sum_{s=1}^{k} |\hat{a}_k(s) - a_k(s)| \to 0$, in probability, uniformly in $1 \leq k \leq K_T$; $|\beta_k(1) - b(1)| = |a_k(1) - a(1)| \to 0$, uniformly in $k_T \leq k \leq K_T$. □

A bound for the h-step mean squared error of prediction of the plug-in method is given in the following two results which for $h = 1$ agree with Shibata (1980) but hold without requiring that k_T, or $\tilde{k}_{PT}$, is a divergent sequence.

PROPOSITION 6.2. *If Assumptions* 1, 2, 3 *with* $s_o = 16$, *and* 4 *hold,*

$$|[\{\|\hat{\boldsymbol{\Phi}}_{Ph}(k) - \boldsymbol{\Phi}_h\|_{\boldsymbol{R}}^2 / F_{PhT}(k)\} - 1]|$$

converges to 0 *in probability, uniformly in* $k_T \leq k \leq K_T$, *for any divergent sequence* $\{k_T\}$ *and each fixed* $h \geq 1$.

PROOF. For any divergent sequence, $\{k_T\}$, as $T \to \infty$,

$$|[\|\hat{\boldsymbol{\Phi}}_{Ph}(k) - \boldsymbol{\Phi}_{Ph}(k)\|_{\boldsymbol{R}(k)}^2 - \{kV(h)/\tilde{N}\}]/F_{PhT}(k)|$$

converges to 0 in probability, uniformly in $k_T \leq k \leq K_T$, by (6.14) and Lemmas 6.1–6.5; also, since, $\{V_P(h,k) - V(h)\} \to 0$, as $k \to \infty$, uniformly in $k_T \leq k \leq K_T$, so does $|w(k)/F_{PhT}(k)|$. □

THEOREM 6.2. *Let Assumptions* 1, 2, 3 *with* $s_o = 16$, *and* 4 *hold. For any random variable,* $\tilde{k}_{PT}$, *possibly dependent on* $x_1, \ldots, x_T$, *and such that* $\tilde{k}_{PT} \to \infty$, *as* $T \to \infty$, *for any* $\delta > 0$ *and each* $h \geq 1$,

$$\lim_{T\to\infty} P([\|\hat{\boldsymbol{\Phi}}_{PhT}(\tilde{k}_{PT}) - \boldsymbol{\Phi}_h\|_{\boldsymbol{R}}^2 / F_{PhT}\{k_{PT}^*(h)\}] \geq 1 - \delta) = 1.$$

PROOF. The result follows from Proposition 6.1, Definition 3.1 and the assumption that $\tilde{k}_{PT} \to \infty$, as $T \to \infty$. □

Theorem 6.2 implies that when k is a random variable, the substitute h-step mean squared error of prediction up to $V(h)$ of the plug-in method, $Q_P(k)$, is never below $F_{PhT}\{k_{PT}^*(h)\}$, in probability. For $h = 1$, as discussed earlier, the bound is attained asymptotically if k is selected by AIC, or an equivalent criterion.

If, however, the same value of k is used for multistep prediction, the actual mean squared error may be greater than the bound; an example is given in Section 7.

A second difficulty with the plug-in method is that for each h, $F_{PhT}\{k^*_{PT}\} \geq F_{DhT}\{k^*_{DT}\}$, the asymptotically attainable bound for the direct method. Two qualifications apply, however, to this comparison. First, the results are asymptotic, and with a finite T, the variance term for the direct method can be expected to be larger than that for the plug-in method; see Bhansali (1993). Second, for an ARMA model, the difference between their bias terms need not necessarily be large, an example is given in Section 7.

7. Examples and extensions

Suppose that $\{x_t\}$ is a moving average process of order 1,

$$x_t = \epsilon_t - \beta\epsilon_{t-1}, \qquad |\beta| < 1, \tag{7.1}$$

where $\{\epsilon_t\}$ is as in Assumption 1.

We have, $a_k(j) = \beta^j(1-\beta^{2k-2j+2})/(1-\beta^{2k+2})$, $j = 1,\ldots,k$, $\sigma^2(k) = \sigma^2(1-\beta^{2k+4})/(1-\beta^{2k+2})$ see Whittle (1963). Hence, for one-step prediction, with $k = 1,\ldots,K_T$,

$$F_{D1T}(k) = F_{P1T}(k) = [\sigma^2\beta^{2k+2}(1-\beta^2)/(1-\beta^{2k+2})] + (k/N)\sigma^2.$$

If $h = 2$, we have, for the direct method, with all $k \geq 1$,

$$\begin{aligned}
&\phi_{Dhk}(j) \equiv 0 \qquad (j = 1,\ldots,k),\\
&V_D(2,k) = V(2) = R(0) = \sigma^2(1+\beta^2),\\
&F_{D2T}(k) = (k/N)R(0);
\end{aligned}$$

and from (6.4) for the plug-in method, with $\beta_k(1) = -a_k(1)$,

$$V_P(2,k) = [1 + \{\beta_k(1)\}^2]\sigma^2(k) + 2\beta_k(1)\{a_k(k)\}^2R(1).$$

Hence, for $h = 2$,

$$F_{PhT}(k) - F_{DhT}(k) = \sigma^2\beta^{2k+2}(1-\beta^2)^2(1-\beta^{2k})/(1-\beta^{2k+2})^3 > 0.$$

If the predictive efficiency of the direct method is judged by the ratio $e(h,k) = V_D(h,k)/V_P(h,k)$, $h > 1$, we have

$$e(2,k) = 1 - O(\beta^{2k+2})$$

converges exponentially to 1 as $k \to \infty$. For $k = 1$,

$$e(2,1) = (1+\beta^2)^4/\{\beta^4 + (1+\beta^2)^4\}.$$

It is seen that, $e(2,1) = 0.975$, if $\beta = 0.5$; $e(2,1) = 0.946$, if $\beta = 0.8$, and the predictive efficiency gain of the direct method is not necessarily substantial.

A related question concerns whether or not the same value of k minimises $F_{PhT}(k)$ for different values of h. That this is not so may be gleaned by a comparison of $F_{P1T}(k)$ and $F_{P2T}(k)$, which shows that the values of k minimising the former and the latter do not always coincide. To illustrate this point, we computed the values of these two functions for $T = 100, 200, 500, 1000$, $K_T = 30$ and several different value of β. For each (T, β) combination, the values of k that minimised $F_{PhT}(k)$, $h = 1$ and 2 were determined. To save space, detailed results are omitted, but suffice to say that for $\beta < 0.5$, the same value of k was selected with $h = 1$ and $h = 2$ and for all choices of T. For $\beta \geq 0.8$, however, there was a significant discrepancy between the two orders selected. For example, with $T = 1000$ and $\beta = 0.8$, the orders selected with $(h = 1, h = 2)$ were $(10, 7)$, which increased to $(17, 8)$, for $\beta = 0.9$ and to $(23, 8)$, for $\beta = 0.95$.

One reason why in the moving average example above the gain in predictive efficiency due to the direct method appears small is that the autoregressive coefficients, $a(j)$, decrease to 0 at an exponential rate as $j \to \infty$; the difference between $V_P(h, k)$ and $V_D(h, k)$ then vanishes extremely rapidly as k increases and even for moderately small values of k, the numerical value of this difference is negligible. It may be anticipated that this would be so for all "short-memory" time series models since for this class of models, by definition, the difference between $V_P(h, k)$ and $V_D(h, k)$ vanishes exponentially as $k \to \infty$.

Stoica and Nehorai (1989) consider multistep prediction of observations simulated from two linear and two non-linear models by fitting an autoregressive model to the simulated data, but with the coefficients estimated by minimising a linear combination of the squared h-step prediction errors, (2.14), $h = 1, \ldots, S$, the maximum lead time. They report that if the second-order properties of the simulated model are well approximated by the fitted autoregressive model then the predictive efficiency gain by using this 'multistep error' method is small in comparison with the usual plug-in method; if, however, an underparametrized model is used for prediction then the gain in predictive efficiency can be considerable.

We note that this finding accords with the discussion given above since in the former situation k may be expected to be sufficiently large to ensure that the difference between $V_P(h, k)$ and $V_D(h, k)$ is small, but the converse is likely to hold in the latter situation when for small values of k the difference can be substantially in favour of the direct method. It should be pointed out nevertheless that the 'multistep error' method used by these authors is not the same as the direct method examined in this paper, since separate model selection for each lead time has not been considered, rather the same model is used at each lead time but with reestimated parameters and the parameter estimation technique employed is also somewhat different from that described in Section 2. In our view, under the hypothesis that the model generating an observed time series is unknown, the question of model selection for each lead time cannot be separated from that of estimation of the multistep prediction constants and they should be considered together, since the forecasts may be improved by doing both simultaneously instead of only the latter.

An application of the autoregressive model fitting approach for prediction of a 'long-memory' time series has been given by Ray (1993) in whose simulations a

fractional noise process is forecast up to 20 steps ahead by the plug-in method, but allowing k to vary over a grid of values without recourse to an order determining criterion. Her results endorse the use of autoregressive models for predicting a long-memory process, but indicate a need for considering models of different orders at different lead times, as envisaged in the direct method.

To illustrate the use of these two methods for multistep forecasting of an actual time series, we consider the first 264 observations of the Wolf Annual Sunspot Series, 1700–1963, as given by Wei ((1990), pp. 446–447). Priestley (1981) observes that this series is unlikely to have been generated by a linear mechanism, and, therefore, the present example should point to how the methods might behave with possibly non-linear data.

The plug-in method was applied by fitting an initial 9th order autoregression, the model selected by the FPE_2 criterion of Bhansali and Downham (1977), and the direct method by the $S_h(k)$ criterion, (2.17)—the orders of the different autoregressive models selected by this criterion for each $h = 1, 2, \ldots, 20$ were 9, 9, 17, 17, 16, 15, 14, 13, 12, 11, 2, 3, 3, 3, 7, 6, 11, 10, 9, and 8, respectively. Thus, for this series, the autoregressive order selected varies considerably with the lead time, h, indicating, first, that the series is highly predictable from its own past—even up to twenty steps ahead, and, secondly, that a finite-order autoregressive process probably does not provide an adequate representation for the generating structure of the series. An explanation for why the selected autoregressive order changes so much with different h is, however, not readily given, but it could be related to the highly periodic and possibly non-linear mechanism generating the data. It may be noted also that the estimates (2.15)–(2.16) do not use the same set of observations for different h; nevertheless, under the stationarity assumption, and with $T = 264$ observations in total, the possible effects of missing out a small number of 'recent' observations may not be critical.

The forecasts obtained by the plug-in and the direct methods of the Sunspot series for 1964–1983 were compared with the recorded values.

As a measure of the forecast accuracy, we use the mean absolute percentage error,

$$(7.2) \qquad MAPE(j) = \left[(20)^{-1} \sum_{h=1}^{20} |Y(h) - \hat{Y}_j(h)| \{Y(h)\}^{-1} \right] \times 100 \qquad (j = 1, 2),$$

where $\hat{Y}_1(h)$ denotes the h-step forecast obtained by the direct method for the year $1963 + h$ and $\hat{Y}_2(h)$ that by the plug-in method and $Y(h)$ denotes the observed value for that year.

The actual numerical values of the metric, (7.2), were computed for both these methods and are shown below:

Direct Method:	21.9%;
Plug-In Method:	23.4%.

The direct method is seen to have a slightly smaller mean absolute percentage error in this example. Nonetheless, for the empirical finding to conform with

the earlier asymptotic results, this difference should be statistically significant, a question we do not pursue here.

The $S_h(k)$, AICh and FPEh criteria may be viewed, see Bhansali and Downham (1977), as special cases with $\alpha = 2$ of the following extended criteria in which $\alpha > 1$ denotes a fixed constant:

$$(7.3) \qquad \begin{aligned} &S_h^{(\alpha)}(k) = (N + \alpha k)\hat{V}_{Dh}(k); \qquad FPEh_\alpha(k) = \hat{V}_{Dh}(k)(1 + \alpha k/N); \\ &AICh_\alpha(k) = T \ln \hat{V}_{Dh}(k) + \alpha k. \end{aligned}$$

Let $\hat{k}_{DT}^{(\alpha)}(h)$ denote the order selected by minimising any of the extended criteria in (7.3). As in Shibata (1980), if the $a(j)$ decrease to 0 exponentially as $j \to \infty$, the selected order would still be asymptotically efficient in the sense of Definition 4.1, but it would not be if the $a(j)$ decrease to 0 geometrically.

Bhansali (1986) defines, with $h = 1$, a class of generalized penalty functions for which the criteria (7.3) are asymptotically efficient with each fixed α. The argument given there extends to $h > 1$ by defining a generalized risk function,

$$F_{DhT}^{(\alpha')}(k) = \{V_D(h,k) - V(k)\} + \alpha'(k/N)V(h),$$

in which $\alpha' > 0$ is a fixed constant. It readily follows from Theorems 4.1 and 5.1 that with $\alpha = \alpha' + 1$ and each $h \geq 1$ the order selected by any of the criteria (7.3) is asymptotically efficient with respect to this extended risk function, provided $\alpha > 1$ remains fixed as $T \to \infty$ and Assumptions 1, 2, 3 with $s_o = 16$ and 4 hold.

Acknowledgements

Thanks are due to Dr. David Findley for drawing the problem considered in the paper to the author's attention, and to the referees and an Associate Editor for their helpful comments.

References

Akaike, H. (1970). Statistical predictor identification, *Ann. Inst. Statist. Math.*, **22**, 203–217.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *2nd International Symposium on Information Theory* (eds. B. N. Petrov and F. Csaki), 278–291, Akademia Kiado, Budapest.

Anderson, T. W. (1971). *The Statistical Analysis of Time Series*, Wiley, New York.

Baxter, G. (1963). A norm inequality for a "finite-section" Wiener-Hopf equation, *Illinois J. Math.*, **7**, 97–103.

Bhansali, R. J. (1981). Effects of not knowing the order of an autoregressive process on the mean squared error of prediction—I, *J. Amer. Statist. Assoc.*, **76**, 588–597.

Bhansali, R. J. (1986). Asymptotically efficient selection of the order by the criterion autoregressive transfer function, *Ann. Statist.*, **14**, 315–325.

Bhansali, R. J. (1993). Estimation of the prediction error variance and an R^2 measure by autoregressive model fitting, *J. Time Ser. Anal.*, **14**, 125–146.

Bhansali, R. J. and Downham, D. Y. (1977). Some properties of the order of an autoregressive model selected by a generalization of Akaike's FPE criterion, *Biometrika*, **64**, 547–551.

Bhansali, R. J. and Papangelou, F. (1991). Convergence of moments of least-squares estimators for the coefficients of an autoregressive process of unknown order, *Ann. Statist.*, **19**, 1155–1162.

Box, G. E. P. and Jenkins, G. M. (1970). *Time Series Analysis: Forecasting and Control*, Holden Day, San Francisco.

Brillinger, D. R. (1975). *Time Series: Data Analysis and Theory*, Holt, New York.

Cox, D. R. (1961). Prediction by exponentially weighted moving averages and related methods, *J. Roy. Statist. Soc. Ser. B*, **23**, 414–422.

del Pino, G. E. and Marshall, P. (1986). Modelling errors in time series and k-step prediction, *Recent Advances in Systems Modelling* (ed. L. Contesse), 87–98, Springer, Berlin.

Findley, D. F. (1983). On the use of multiple models for multi-period forecasting, *Proceedings of the Business and Economic Statistics Section*, 528–531, American Statistical Association, Washington, D.C.

Gersch, W. and Kitagawa, G. (1983). The prediction of a time series with trends and seasonalities, *Journal of Business and Economic Statistics*, **1**, 253–264.

Greco, C., Menga, G., Mosca, E. and Zappa, G. (1984). Performance improvements of self-tuning controllers by multistep horizons: the MUSMAR approach, *Automatica*, **20**, 681–699.

Grenander, U. and Rosenblatt, M. (1954). An extension of a theorem of G. Szego and its application to the study of stochastic processes, *Trans. Amer. Math. Soc.*, **76**, 112–126.

Hurvich, C. M. (1987). Automatic selection of a linear predictor through frequency domain cross validation, *Comm. Statist. Theory Methods*, **16**, 3199–3234.

Kabaila, P. V. (1981). Estimation based on one step ahead prediction versus estimation based on multi-step ahead prediction, *Stochastics*, **6**, 43–55.

Lewis, R. and Reinsel, G. C. (1985). Prediction of multivariate time series by autoregressive model fitting, *J. Multivariate Anal.*, **16**, 393–411.

Lin, J.-L. and Granger, C. W. J. (1994). Forecasting from non-linear models in practice, *Journal of Forecasting*, **13**, 1–9.

Milanese, M. and Tempo, R. (1985). Optimal algorithms theory for robust estimation and prediction, *IEEE Trans. Automat. Control*, **AC-30**, 730–738.

Pillai, S. U., Shim, T. I. and Benteftifa, M. (1992). A new spectrum extension method that maximizes the multistep minimum prediction error-generalization of the maximum entropy concept, *IEEE Transactions on Signal Processing*, **40**, 142–158.

Priestley, M. B. (1981). *Spectral Analysis and Time Series*, Academic Press, New York.

Ray, B. K. (1993). Modeling long-memory processes for optimal long-range prediction, *J. Time Ser. Anal.*, **14**, 511–526.

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process, *Ann. Statist.*, **8**, 147–164.

Shibata, R. (1981). An optimal autoregressive spectral estimate, *Ann. Statist.*, **9**, 300–306.

Stoica, P. and Nehorai, A. (1989). On multistep prediction error methods for time series models, *Journal of Forecasting*, **8**, 357–368.

Stoica, P. and Soderstrom, T. (1984). Uniqueness of estimated k-step prediction models of ARMA processes, *Systems Control Lett.*, **4**, 325–331.

Tiao, G. C. and Xu, D. (1993). Robustness of MLE for multi-step predictions: the exponential smoothing case, *Biometrika*, **80**, 623–641.

Wall, K. D. and Correia, C. (1989). A preference-based method for forecasting combination, *Journal of Forecasting*, **8**, 269–292.

Wei, W. W. S. (1990). *Time Series Analysis*, Addison Wesley, Redwood City.

Weiss, A. A. (1991). Multi-step estimation and forecasting in dynamic models, *J. Econometrics*, **48**, 135–149.

Whittle, P. (1963). *Prediction and Regulation by Linear Least-Square Methods*, English University Press, London.

Yamamoto, T. (1976). Asymptotic mean square prediction error for an autoregressive model with estimated coefficients, *Applied Statistics*, **25**, 123–127.